ARMY RESEARCH LABORATORY

Selection of a Workload Metric for Evaluation of Telemedicine Applications: Literature Review and Methodological Development

Leslie A. Whitaker
Jennifer Hahus
Deborah Birkmire-Peters

ARL-TR-1264 MAY 1997

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

ARL-TR-1264 May 1997

Selection of a Workload Metric for Evaluation of Telemedicine Applications: Literature Review and Methodological Development

Leslie A. Whitaker
Jennifer Hahus
University of Dayton

Deborah Birkmire-Peters
Human Research & Engineering Directorate

Abstract

A measure of cognitive workload was needed to conduct human factors evaluations of telemedicine applications. A literature review was conducted to find available metrics and select candidates for testing. Three candidate measures (the Subjective Workload Assessment Technique [SWAT], NASA-Task Load Index [TLX] along with its subscales, and the Modified Copper-Harper [MCH]) were selected using the following criteria: reliability, validity, lack of contamination, availability, sensitivity, lack of intrusiveness, diagnosticity, and cost. All metrics in the literature review, as well as the application of the selection criteria, are described in this report. Methodological development and research were then completed to develop a research paradigm for selecting the best workload metric from the three candidates. This effort included the development and norming of difficulty levels of a surrogate task in a controlled experimental protocol, the selection of a spatial abilities test, acquisition and testing of required telecommunication and recording equipment, and the iterative development and testing of a research protocol. These processes and their results are described in detail in this report.

CONTENTS

SELECTION OF A WORKLOAD METRIC FOR EVALUATION OF TELEMEDICINE APPLICATIONS: LITERATURE REVIEW AND METHODOLOGICAL DEVELOPMENT

INTRODUCTION

Telemedicine literally means medicine at a distance. Presently, telemedicine has been defined as the use of telecommunications and information technologies to provide health care. This encompasses the diagnosis, treatment, monitoring, and education of patients regardless of the patient, provider, or information location (Puskin, Brink, Mintzer, & Wasem, 1995).

There have been a number of efforts to use telemedicine to deliver health care to remote and medically underserved populations over the last 40 years. A review of the telemedicine programs during this time, however, revealed that only one major project continued to survive after the withdrawal of external funding (Hassel, 1995). The reasons for the lack of success of these telemedicine efforts are not apparent. This is in large part because few, if any, rigorous scientific evaluations were done.

The problem of evaluating telemedicine applications has recently been recognized and addressed by a number of researchers and policy makers in the area (Bashshur, 1995; Grigsby, Schlenker, Kaehny, Shaughnessy, & Sandberg, 1995; Puskin et al., 1995). In particular, the Department of Defense (DoD) Telemedicine Evaluation Working Group (TEWG) proposed a conceptual framework to guide the development of methodologies to evaluate telemedicine projects in the DoD. The five areas to be evaluated in the TEWG framework are clinical outcomes, patient-provider satisfaction, human factors, organizational impact, and costs and benefits.

One of the areas in the human factors evaluation that was determined to be important was the assessment of workload. It has been shown in other areas (e.g., aviation) that changes in technological applications have resulted in additional workload demands on the operator. This additional workload has been related to decrements in performance. It is believed that a similar change in the behavioral and cognitive workload of the health care provider may occur as a result of the additional requirements imposed by telemedicine applications. This change in workload may result in an increase in the number of errors committed.

Consequently, a review of the cognitive workload literature was completed to identify the most promising workload metrics for possible use in measuring changes in workload in

telemedicine applications. The literature review and the selection of the candidate measures is described in the next section of this report.

It is necessary to subject these candidate measures to empirical verification and validation for use in evaluating telemedicine applications. To maintain experimental control, as well as for legal and ethical reasons, the original verification and validation process must be conducted in a laboratory before being used in evaluating actual telemedicine applications. Therefore, a surrogate laboratory task, which taps the same cognitive demands as expected in telemedicine applications, and a laboratory protocol for testing workload metrics were developed. This task and protocol development are described in the present paper.

The work reported here will serve as the basis for further development of a methodology for evaluating workload in telemedicine applications. The potential metrics need to be verified and validated with more appropriate populations. Following that work, it should be possible to extend the findings of this research to the evaluation of actual telemedicine applications.

LITERATURE REVIEW

Database

A detailed analysis of subjective workload metrics was used to select the metrics that hold the most promise for use in telemedicine. This analysis examined human factors technical and psychology electronic databases using the terms *workload and subjective; cognitive workload and subjective; mental workload and subjective*; plus several specific metric names--*overall workload; OW; SWAT; NASA-TLX; TLX; Cooper-Harper; MCH.* From the titles accessed by this search, three comprehensive reviews, three meta-analyses, and 44 articles describing experimental results are selected as a comprehensive information set upon which to base our metric selection decisions.

Workload Measurement Classification

Cognitive workload measures can be classified into three broad areas: physiological, performance (primary task or loading task), and subjective (Schlegel, 1993).

Physiological Measures

The human body responds both cognitively and physiologically to the demands of

its environment and its tasks. Physiological measures that vary with cognitive demands have been tested as potential metrics of those cognitive demands (Wierwille & Eggemeier, 1993). These measures include eye blink rate, pupil diameter, P300 amplitude and latency, galvanic skin response (GSR), heart rate, heart rate variability, and certain blood and urine fractions (e.g., norepinephrine). It is difficult to measure physiological responses because of the large number of trials that must be performed to obtain reliable measures and because of the invasive nature of most of the measurement technology. Finally, those measures that have been obtained often do not agree with one another (i.e., a task demand may be reflected in heart rate variability but not in GSR) and do not consistently occur in the literature.

Performance Measures

Performance measures for the primary task are the most direct indication of changing cognitive workload (Crabtree, Bateman, & Acton, 1984). When a task requires primarily cognitive effort, changes in that task's performance might be thought to provide the best indication of changes in the level of cognitive effort. However, this will only prove to be the case if the performance is sensitive to these changes in workload (Boff & Lincoln, 1988). Instead, suppose that a person can perform a task with low workload demands without error and without employing the maximum cognitive resources to complete the task. Then suppose that the demands of the task are increased; now the worker can continue to perform the task without error only by expending all his or her cognitive resources; that is, he or she has no spare resources but is able to maintain flawless performance. In this way, performance is not a sensitive indicator of the changes in cognitive workload.

A reasonable question to ask is why one would care about changes in workload that do not affect the performance of the task of interest. When a task is completed during testing conditions, we usually find that the operator is rested, the communication among team members is perfect, the time on task was limited, and no emergencies occurred. In these circumstances, task performance may not be a sensitive indicator of how close an operator is to using all available resources. However, whenever any one of these circumstances is compromised, as they often are during actual operating conditions, then the operator using all his or her cognitive resources to maintain flawless performance during optimal circumstances will be overloaded and begin to make errors. In contrast, the operator completing a task with a lower workload will have an available cognitive reserve to muster in the face of adverse circumstances. This is the reason that a sensitive measure of workload may provide a better predictor of operational performance than could tested performance itself.

One means of improving the sensitivity of performance measures is to add an additional task that will use all available cognitive resources even during normal testing conditions (Fisk, Derrick, & Schneider, 1983). This procedure requires that the operator complete two tasks concurrently; one is the task of interest (the primary task) and the second is a loading task used to push the demands on the operator's resources, even during the lightest primary task workload conditions. This is known as a dual task paradigm. The result is that operator performance of the combination of tasks demands all cognitive resources at each level of primary task workload. In this way, changes in that workload will be accurately reflected in changes in performance of one or both of the concurrent tasks. In effect, the loading task is acting in much the same way that the adverse circumstances and emergency demands of the operational setting affect cognitive demands and in turn, adversely affect task performance.

Subjective Measures

Operators are capable of describing the difficulty of a task. Various measurement instruments have been designed to quantify these difficulty evaluations (Gopher & Donchin, 1986). These are known as subjective measures of workload. Because the cognitive workload involved in the completion of many tasks is the conscious work that occurs in working memory, that is, short term memory, the operators themselves are able to report the amount of cognitive effort expended. Hence, numerous publications over the past 20 years have reported the effectiveness of subjective workload metrics in assessing cognitive workload. In addition to their sensitivity and implied reliability, these measures have face validity and have provided validity when compared with task performance (Eggemeier, McGhee, & Reid, 1983; Boyd, 1983). They are relatively inexpensive to collect and are usually nonintrusive on the task itself. That is, the subjective workload measure can be collected without interfering with task performance (Eggemeier, Melville, & Crabtree, 1984).

Criteria for Selection

The goal for the present review is to determine candidate workload measures for the assessment of cognitive workload in telemedicine applications. To be useful, any measurement must meet four criteria: *reliability, validity, lack of contamination,* and *availability.* Successful workload metrics should meet four additional (and not strictly independent) criteria: *sensitive, nonintrusive, diagnosticity,* and *cost effectiveness.* Therefore, each candidate class of measures and each measure itself will be evaluated for these criteria.

The criteria are defined and described in the following section:

1. Reliability is the repeatability of a measure. When a measure is reliable, then repeated occasions, similar tasks, or judges will obtain similar measurement levels. Without reliability, a measure cannot be sensitive or valid. Therefore, finding validity and sensitivity implies that reliability exists; however, it is far better to assess reliability directly, although this is too seldom done in operational settings (Lysaght et al., 1989).

2. Validity is the degree to which a metric actually measures the concept it is intended to measure. For example, an intelligence test is valid if it measures abilities as opposed to measuring achievement.

3. Contamination occurs when a metric is confounded with other influences, unrelated to the measurement of interest. For example, contamination in workload measures would occur when physical effort to complete the workload assessment confounds the measurement of cognitive workload for the task per se. Lack of contamination is important to any satisfactory metric.

4. Availability indicates the ability to obtain the measurement. Availability may be limited by access, funding, or intrusiveness into the task domain itself.

5. Sensitivity is the extent to which changes in the item to be measured are reflected by changes in the measuring instrument. Lack of sensitivity will decrease both reliability and validity. An example of an insensitive workload measure was given earlier in the form of some primary task performance measurements.

6. Intrusiveness means the extent to which performance of the primary task is interrupted by the workload metric. Any concurrent demands for obtaining the measurement of workload have the potential to intrude on the primary task, but not all appear to do so. Nonintrusiveness is an important criterion of a useful workload metric.

7. Diagnosticity refers to the ability of a metric to determine what aspect of the task is the source of the imposed workload, that is, what operator resource is more severely taxed (see Polzella & Reid, 1987, and Vidulich & Wickens, 1986, for contrasting views). If an unacceptably high workload is found, then a diagnostic metric will pinpoint the cause of that overload.

8. Cost must be evaluated against the value obtained from knowing the workload information. The relationship between the value of the workload information obtained and the cost of obtaining it is the cost effectiveness of the metric.

Application of Criteria to Workload Metrics

These evaluation criteria can be applied to each of the three broad classifications of workload metrics: physiological, performance (primary and dual), and subjective.

Physiological measures have lacked reliability during similar test-retest conditions. Furthermore, when multiple physiological measures are obtained, they often do not correlate with one another in reflecting changes in cognitive workload. Without reliability, validity is not possible; therefore, the question of validity can only be considered when a physiological measure has been reliable. Physiological measures are frequently contaminated by artifacts from other physiological activities (e.g., eye blinks, breathing, or muscle movements). Although some physiological measures can be obtained directly, most interest in the assessment of cognitive workload (e.g., P300 evoked brain potentials) requires the use of high technology equipment to measure small electrical impulses, separate them from surrounding signals, and analyze them statistically. The sensitivity of these measures has been found in some cases, but often it is not found. A specific application of P300 in the measurement of perceptual workload has been found when using a secondary task to elicit the P300. In this case, some diagnosticity was found (Gopher & Donchin, 1986). Finally, the need for equipment attached to the operator results in very intrusive and expensive measurement methodology.

Performance measures might be thought to be reliable and valid measures of cognitive workload solely by their definition. This statement assumes that performance results from cognitive workload alone. However, especially when using only a primary task, this has not always been the case. Employing a second loading task has improved the sensitivity of performance as an indicator of cognitive workload. Unfortunately, the use of dual task paradigms may result in decrements in the primary task or the loading task or both, as workload increases. This may compromise the safety of the primary task in an operational setting, and even in an experimental setting, it makes interpretation of the results difficult. The only sources of contamination that have been reported are the cross linking of demands from the two concurrent tasks. Intrusion from the loading task can be alleviated by careful selection of the loading task itself. One successful method has been to develop imbedded secondary tasks

specific to each type of primary task being evaluated. The costs of obtaining performance measures, whether primary or loading task performance, are moderate.

Reliable subjective measures have been developed (e.g., SWAT and NASA-TLX). This cannot be claimed for all subjective workload measures that have been employed (Gopher & Donchin, 1986). Furthermore, cluster analyses (Derrick, 1983) have confirmed that these measures are valid in assessing a variety of the cognitive demands that impact workload. These measures can be easily contaminated by experimenter expectations and operator motivation. Care must be taken to avoid these problems when using subjective workload measures, and the procedures for administering the well-developed metrics have taken these precautions. Standard metrics for assessing subjective workload have been established for other domains such as flight and communication, but they have not been employed in telemedicine applications. The sensitivity of some metrics has accurately reflected changes in cognitive workload demands (e.g., signal rate, short term memory, and auditory communication requirements) (Eggemeier, Crabtree, & LaPointe, 1983; Moroney, Biers, & Eggemeier, 1995). These metrics can be collected after the primary task is completed, and hence, they are nonintrusive; their cost is low. See Table 1 for a summary of this analysis.

In the initial analysis of the three broad classes used to measure cognitive workload, the category of subjective workload metrics is the most satisfactory when evaluated by these test and evaluation criteria. They meet the standards of reliability, validity, and lack of contamination. Several metrics have been standardized and have been tested in other domains. In these domains, such metrics have been sensitive indicators of workload, as well as predictors of task performance. In general, subjective metrics are not thought to be global indicators of workload; they are not particularly diagnostic of the source of this overload. They are the least expensive of all metrics (other than observing primary task performance alone). The nonintrusive nature of subjective workload measures is a very important criterion for their use in the operational settings of telemedicine practices.

Table 1

Evaluation of Broad Workload Classifications

	Workload classification			
		Performance		
Criterion	Physiological	Primary	Loading	Subjective
Reliability	Poor	Good	Good	Generally good
Validity	Variable	Variable	Good	Good
Contamination	Variable	Variable	Variable	Good
Availability	Poor	Good	Variable	Good
Sensitivity	Variable	Variable	Good	Good
Intrusive	Poor	Good	Variable	Good
Diagnostic	Good(P300)	Poor	Good	Poor
Cost	Poor	Good	Moderate	Good

Subjective Workload Metrics

Background

Moray (1982) published a comprehensive review of subjective mental workload examining the literature from 1968, when cognitive measures of performance were first beginning to be examined by the human factors community. He reports that few studies had been published during that time, but his analysis of those studies is particularly helpful for the present task: selecting workload metrics for telemedicine applications. This review was divided into four categories, of which, three are relevant to the cognitive demands of telemedicine procedures: *cognitive, manual control,* and *time stress tasks.*

For the analysis of cognitive tasks, a global measure of subjective workload (such as "On a scale from 1 to 9, how difficult is this task?") correlated better than $r = 0.90$ with task performance. This result has tended to be substantiated by experimental results in the ensuing decade when primary task paradigms were tested; however, global subjective workload measures have been found to dissociate from performance when dual task paradigms or tasks requiring either overlearned (automated) or complex responses are employed (Wickens & Yei-Yu, 1983; Vidulich & Wickens, 1986).

Manual control tasks assessed were all flight control tasks. The primary assessment tool was a subjective rating scale of handling characteristics called the Cooper-Harper (CH) scale. The focus of this review was on the characteristics of the manual control tasks that affected subjective workload. Both order of control and display-to-response lag increased subjective workload. This is consistent with the performance literature which has found increases in error rates with as little lag as 250 msec in speech signals. The upper limit on lag that can be accommodated at all in continuous manual control tasks is 5 seconds. Furthermore, the requirement to complete concurrent manual control tasks and the introduction of instability into the control system also reliably increased workload ratings on the CH scale. A medical analogue to this manual control task occurs in laparoscopic gall bladder surgery when more than one manipulator must be controlled inside a patient's closed abdomen. This lapyroscopic surgery is analogous to teleproctored surgery because the surgeon must view the operation indirectly through a display on a color monitor. If remote transmission produces a lag in the visual display system, a major source of documented workload will be introduced. The CH scale has been sensitive to this lag.

Finally, time stress has been an important driver of cognitive workload and is a factor in some medical procedures considered for telemedicine intervention (e.g., surgery, emergency room medicine). Philipp, Reiche, and Kirchner (1971) found that the workload for air traffic controllers who were on duty for several hours could be assessed using a nine-point scale for two global questions: *How difficult is the task?* and *How much time stress is there?* Objective measures of information processed and time pressure for communication were correlated with the two subjective measures. These correlations were $r = 0.69$ and $r = 0.56$, respectively, indicating a significant relationship between the objective and the subjective measures. These correlation levels are well within the accepted levels for measuring validity.

This background describes the subjective workload research issues that emerged along with a revival of general interest in cognitive psychology approximately 25 years ago. Subsequent interest in this method of assessing workload has resulted in a number of tested subjective workload assessment techniques. These metrics are described next.

Candidate Measurement Tools

A number of candidate metrics have been developed and tested (see Lysaght et al., 1989, and Boff & Lincoln, 1988, for reviews). The present analysis targets metrics that may be of particular use in the assessment of cognitive workload in telemedicine.

Subjective workload metrics may be divided into two general categories: rating scales, which provide quantitative measures of subjective workload, and questionnaires and interviews, which provide qualitative information and lessons learned. Many measures of subjective workload have been developed solely for their application to a single study or to a single area. These measures are not discussed since only measures that have hopes of generalizing from other domains are reasonable candidates for evaluating telemedicine applications. Several ratings scales have been subjected to test and evaluation development and shown to be valid in previous research and will be considered. They are described in the following section:

- **Cooper-Harper Scale (including modified Cooper-Harper)** is a widely used metric which was originally developed for assessing aircraft-handling capabilities. It has been a sensitive indicator of workload for motor or psychomotor tasks (Wierwille & Connor, 1983). A modified version called the modified Cooper-Harper (MCH) has been used successfully to assess perceptual and cognitive requirements (Wierwille & Casali, 1983). One factor to consider in using the CH or MCH is that it is a rating scale which produces only ordinal scale data, thus limiting analysis of statistical significance to non-parametric tests.

- **NASA-Task Load Index (NASA-TLX) and its subscales** is a group of six scales reflecting separate dimensions of workload and an overall workload rating (Hart & Mashkati, 1988). These dimensions include cognitive loading factors such as time pressure and mental effort, as well as physical factors such as amount of physical effort. The rating is a 20-point scale which is assumed to be interval. The NASA-TLX has undergone extensive and rigorous theoretical development and evaluation. Although the TLX has been used most extensively to evaluate flight tasks, it has been used to assess workload in laboratory tasks (e.g., short term memory, visual search, and target acquisition). It has been found to be a valid, reliable, and sensitive measure of cognitive workload. The TLX is preferred to the longer NASA-bipolar measure because of the easier administration of the TLX and the failure to demonstrate an advantage of the bipolar version.

- **Subjective Workload Assessment Technique (SWAT)** is a group of three scales reflecting separate dimensions of workload: time pressure, mental stress, and effort. SWAT has undergone extensive theoretical development and has been

evaluated in both aviation and non-aviation environments (e.g., Eggemeier & Stadler, 1984; Eggleston, 1984; Heffley, 1983; Detro, 1985). Use of conjoint measurement converts these subscale ratings into a single workload measure which is interval instead of ordinal (Nygren, 1991). This metric has been a sensitive, reliable, and valid measure of cognitive workload. As currently used, there are only three rating levels for each dimension (subscale). As the result of current test and evaluation studies (see Moroney, Biers, & Eggemeier, 1995; Biers & McInerney, 1988), it may be possible to eliminate the current scaling procedure necessary for conjoint measurement. This scaling procedure (called a card sort) has limited the number of levels on each subscale. If the card sort is not used, increasing the levels on each subscale from three to five may improve sensitivity and remove floor and ceiling efforts.

- **Psychophysical Scaling (e.g., magnitude estimation)** asks that operators report the workload imposed by a task in comparison to some other task or standard. For example, using *magnitude estimation*, a standard task will be assigned a numerical value and operators are asked to compare a task's workload to that of the standard task by assigning a numerical value to the current task. Using *paired comparisons*, all tasks are paired and the operator chooses the one of the pair with the higher subjective workload (see Acton, Crabtree, Simons, Gomer, & Eckel, 1983, for an application). The difficulty with this procedure is that number of pairs of tasks (n)(n-1)/2 increases too rapidly as the number of tasks themselves (n) increases. Equal-appearing intervals ask operators to assign tasks to categories judged to be of increasing difficulty. The categories are interval scales. Although extensive work has been done in the development of psychophysical scaling techniques for judging laboratory stimuli, little work has been reported from operational settings or from workload measurement. The potential is there, but it awaits further work to determine its applicability.

- **Stockholm Scales** are the result of early work at the University of Stockholm in the development of a univariate (nondimensional) measure of workload. This measure was validated using items on an intelligence test that measured spatial ability, reasoning ability, and verbal comprehension. (Note that all these tasks are processed in conscious or working memory and hence should be readily available for subjective evaluation by the subject.) The reliability and validity as measured by this evaluation were very high. An 11-point version of this scale was used to

assess spare mental capacity in a dual task paradigm using laboratory tasks. These tasks were either perceptually demanding (e.g., target acquisition) or demanding of central processing capacity. In both cases, the Stockholm Scale correlated well with performance and secondary task measures of spare capacity (i.e., it was sensitive to changes in the primary task difficulty). The scale is designed to measure effort as available spare central processing capacity, not motor or psychomotor control.

- **Overall Workload (OW)** Each of the scales described can be used as an overall workload measure; some (e.g., NASA-TLX and SWAT) also have subscales that may allow diagnostic analysis of the source of the workload when overload occurs (Hendy, Hamilton, & Landry, 1993). The initial focus of a workload analysis is to determine whether there is an overload that must be remedied before the task can be completed safely with a reasonable degree of operator workload. If overload is found, further cognitive task analysis can be used to evaluate the cause.

Conclusions

The three measurement scales that have undergone most extensive theoretical development and are most relevant for the present evaluation are the MCH, NASA-TLX, and SWAT. Each has been a valid and reliable predictor of workload in several fields. Of the three, MCH appears to be the most likely to measure any motor or psychomotor components of a medical procedure. NASA-TLX has had less testing outside the aviation world than has SWAT, but it has been shown to correlate well with SWAT and MCH results in those cases when two or more of these metrics have been tested together (e.g., Vidulich & Tsang, 1986; Warr, Colle, & Reid, 1986; see also Lysaght et al., 1989, for a summary review). SWAT has been a sensitive predictor of increasing task difficulty, measuring increased workload before the point that task difficulty leads to a decrement in performance (Whitaker, Peters, & Garinther, 1989).

Each of these metrics has been more or less sensitive to changes in task difficulty, depending upon the domain in which they have been used. This domain-specific aspect requires that comparisons be made among these candidate measures to determine which is the most effective in evaluating cognitive workload for various telemedicine procedures.

The following sections of this report describe the development of a research protocol, which can be used to determine the most sensitive workload metric from among the candidate pool.

METHODOLOGICAL DEVELOPMENT AND EXPERIMENTATION

A performance task using puzzle patterns was developed, and the difficulty level of each pattern was assessed in an experimental protocol. A measure of individual differences was obtained and tested, and a research protocol was developed through testing of the surrogate task and telecommunications equipment. The results of this effort are described in the present report.

Development and Assessment of Performance Task: Puzzle Patterns

Rationale

To maintain experimental control, as well as for ethical and legal reasons, it was not possible to use an actual medical procedure in the planned assessment of workload metrics. Therefore, an alternate task that shared the cognitive demands of such procedures was needed. The following demands were considered to be essential for a surrogate task:

1. Teamwork—Teamwork between at least two team members is required. In a telemedicine application, at least one person is located remotely. He or she is communicating with either another health practitioner or a patient at a distance.

2. Visual-Spatial Requirement—Many telemedicine applications require the transmission of video images to be evaluated by a remotely located specialist. Therefore, the task had to incorporate the visual-spatial requirements of those telemedicine applications.

3. Communication—A communication component was needed because one way in which face-to-face (also called co-located) conditions and telemedicine conditions differ is in the need for one health care practitioner to provide information to a remotely located health care practitioner via audio-video channels.

4. Performance Demands—The task had to place accuracy and time pressure constraints on the team members so that the outcome will produce sensitive

performance indicators. In this way, it is possible to assess the correlation between successful task execution and subjective workload.

5. Psychomotor Component—Many medical procedures have a large psychomotor component. The ultimate goal for the selected surrogate task is that it will be useful for assessing workload during medical procedures. Therefore, a task that has a psychomotor component was needed.

Development of a Surrogate Task

Using this rationale as the basis for selecting a surrogate task, a search for existing normed and validated tasks was conducted. A potential match was found in the spatial abilities block pattern task of the WAIS-R Intelligence Test (Wechsler, 1981). In this test, a person is suppose to construct a two-color pattern from blocks, which matches the pattern shown on a display card. The WAIS-R contains five four-block patterns and four nine-block patterns. This task met each of the four cognitive criteria established for selecting a surrogate task and can be modified to include a psychomotor component. Furthermore, it has validity in that manipulation of blocks to form a pattern is used in the training of surgeons for ophthalmic procedures.

Available Norming Data

The test manual claims that, during the development of the WAIS-R, performance data were obtained to measure the difficulty of the nine patterns. However, these norming data were not available from either the research department or the legal department of Psychological Corporation, despite repeated inquiries. The following information was available: (a) the earlier patterns are easier than the later patterns, and (b) all four-block patterns are easier than all nine-block patterns. Therefore, only ordinal scaling was assumed and the number of difficulty levels was not known.

Creating Additional Patterns

The design of the experimental protocol for the application of this surrogate task was going to require as many as 54 different puzzle patterns. WAIS-R provided only nine patterns. Therefore, it was necessary to develop many additional patterns. These additional patterns were developed in the following ways:

- The original pattern was rotated 90° or 180°. A rotation of 30° is scored as a different pattern in the WAIS-R; therefore, any rotation greater than 30° should be discriminable.
- The colors were reversed.
- A random change was made in one of the original, rotated, or reversed patterns to generate a discriminable pattern with a similar appearance.

Four of the original WAIS-R patterns were used and 11 rotated patterns, five color reversal patterns, and 36 random alteration patterns were added to the original set to produce a complete set of 56 puzzle patterns. Each pattern was assigned a letter code ranging from A through ddd in random order. These 56 patterns are shown in Figure 1.

Assessing Pattern Difficulty

Numerous scaling methods can be used to assess perceived task difficulty. Two have been sensitive, reliable, valid, uncontaminated, and manageable: magnitude estimation and rank ordering (Kling & Riggs, 1972). These two methods and their application to this assessment are described next.

Magnitude estimation asks the observer to assign a number to each item being assessed. This number is to reflect the level of the variable being assessed (in this case, pattern difficulty). A range of possible magnitudes is given and sometimes an anchoring value is used, although this anchor can lead to distortions. Magnitude estimation can produce interval scaled data. In this specific case, an observer was shown a set of cards, each showing one of the 27 four-block patterns. The observer was allowed to look at each of the patterns and to make any comparisons while examining the set. Next, the experimenter shuffled the cards and then showed the cards one at a time and asked the observer to assign a magnitude between 1 and 50 to each card. The 29 nine-block patterns were assessed in the same way except that the range of magnitudes was 51 to 100.

Rank Ordering asks the observer to place the items in an order of increasing value on the variable being assessed. Rank ordering produces ordinal scaled data. In this case, after completing the magnitude estimation task for the four-block patterns, the experimenter again shuffled the cards and then asked the observer to place the cards in order from the easiest to the most difficult pattern. After completing the magnitude estimation task for the nine-block patterns, the observer rank ordered this set.

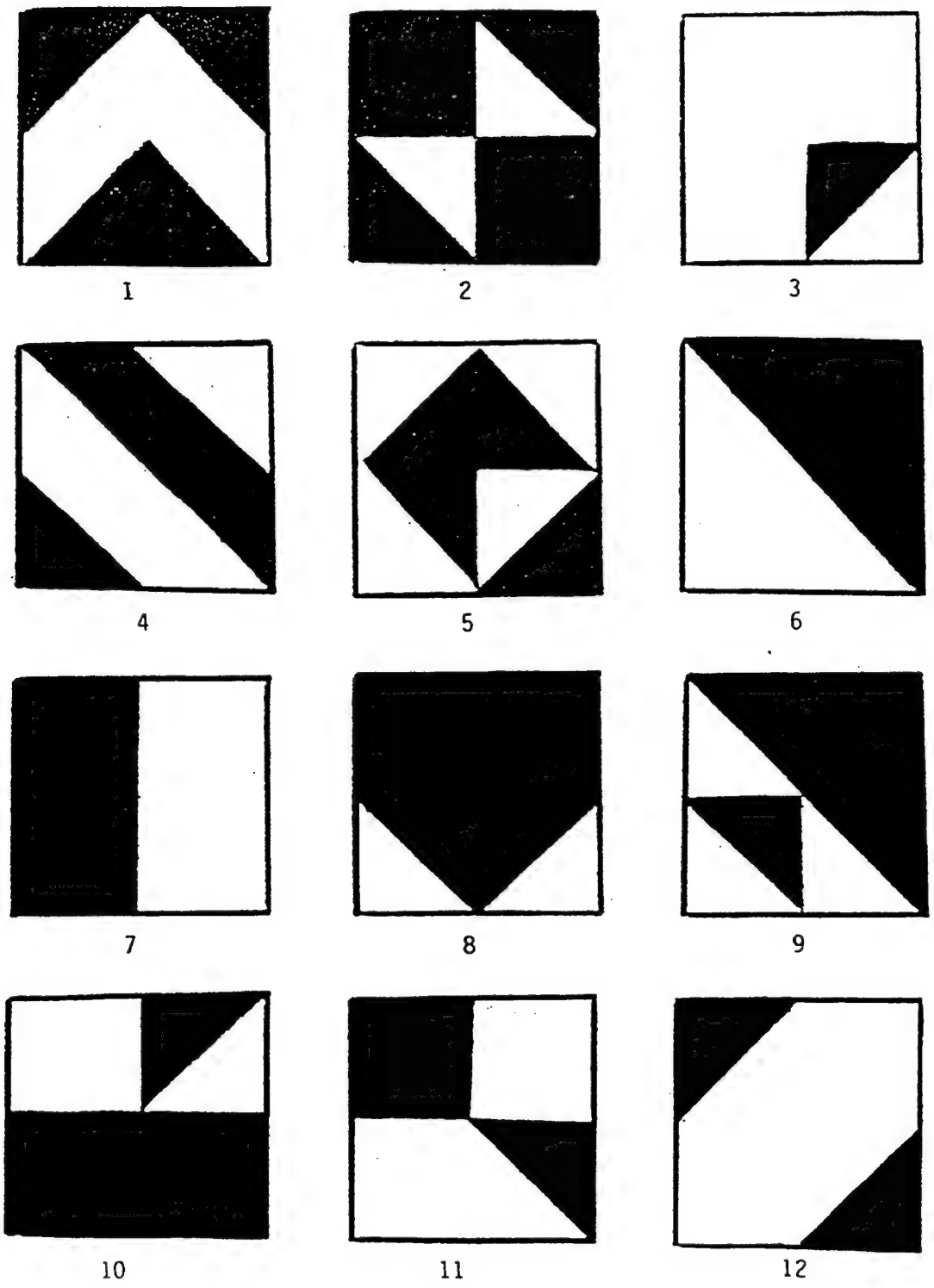


Figure 1. Puzzle patterns developed for experimental paradigm.

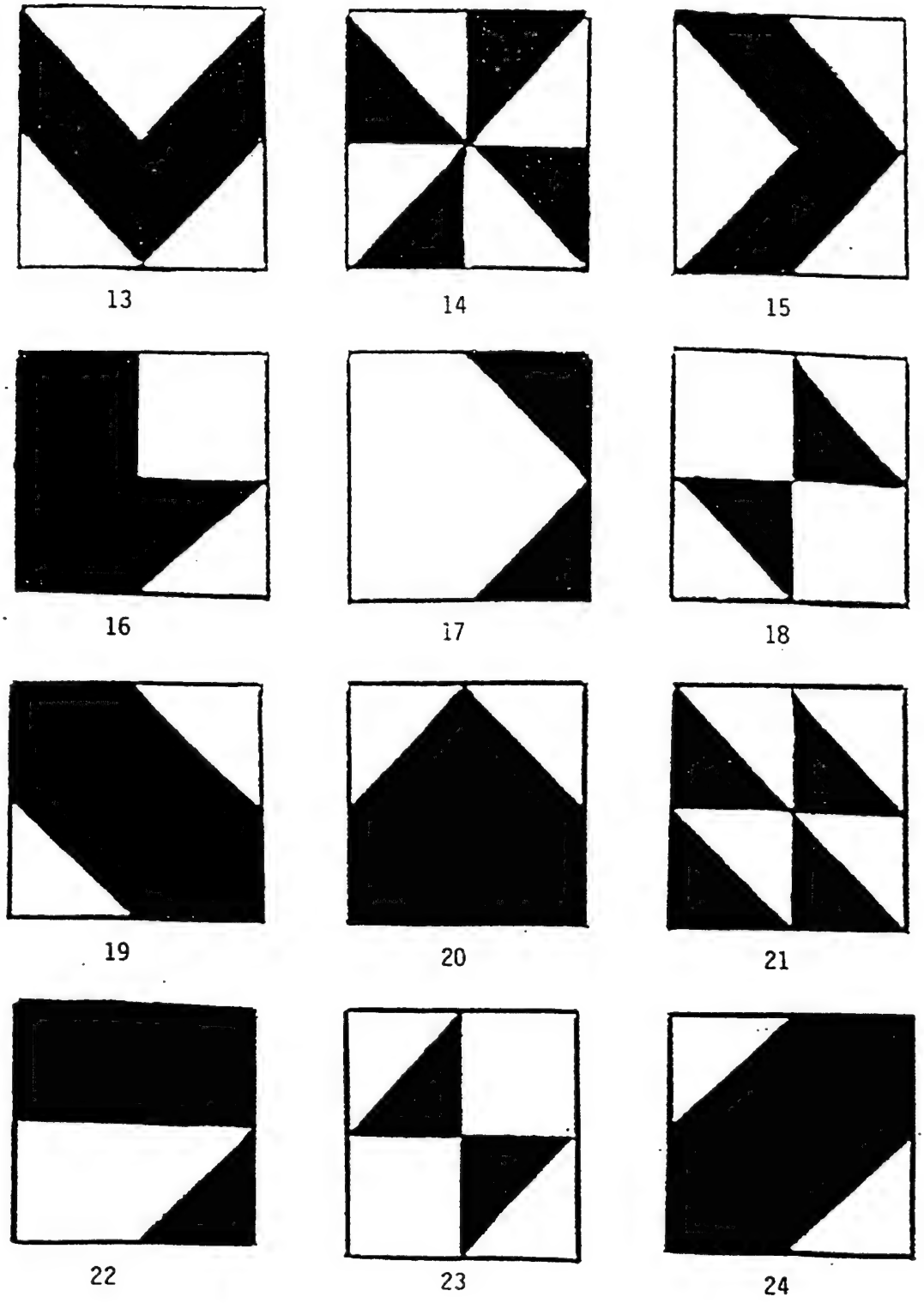


Figure 1. (continued)

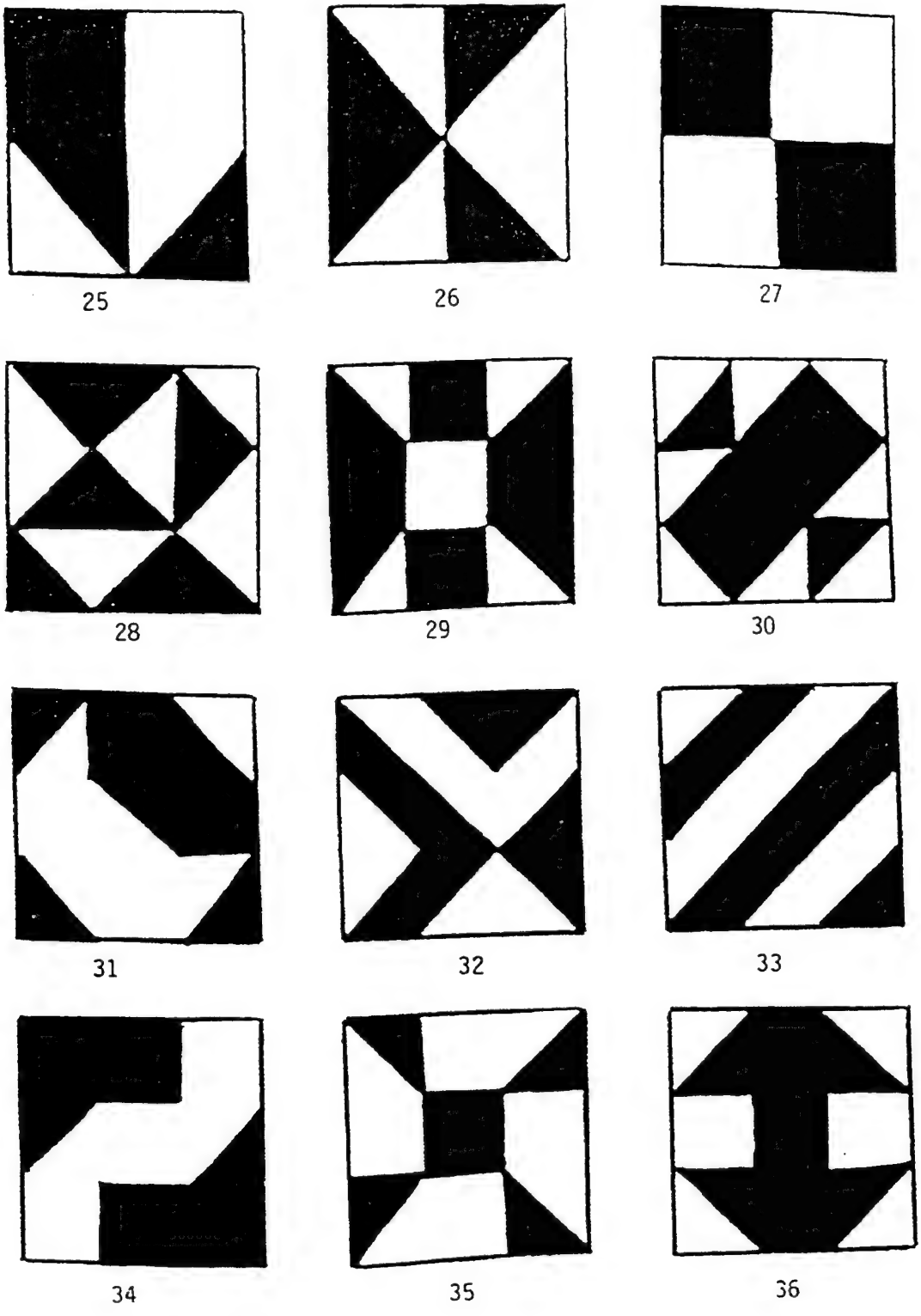


Figure 1. (continued)

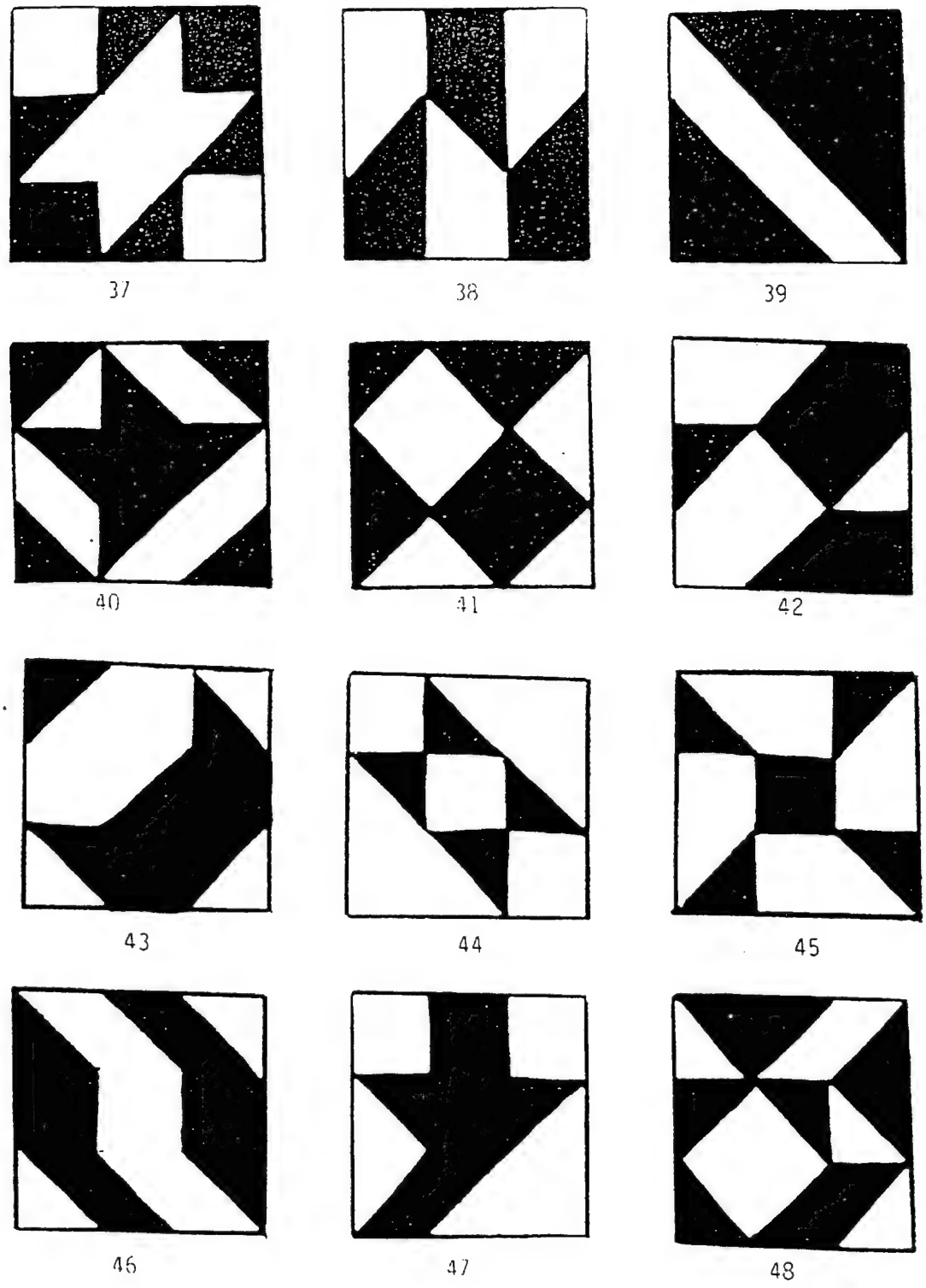
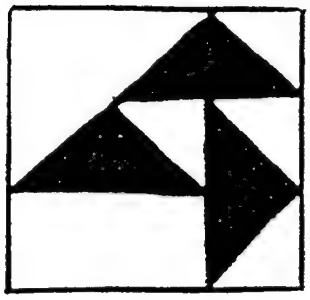
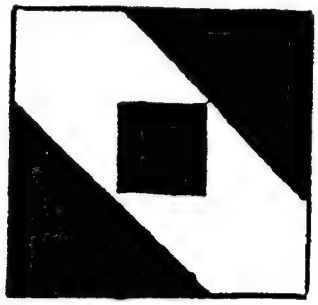


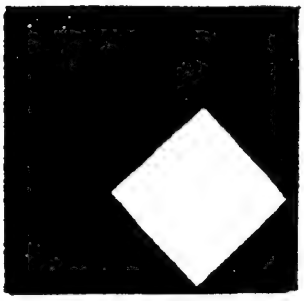
Figure 1. (continued)



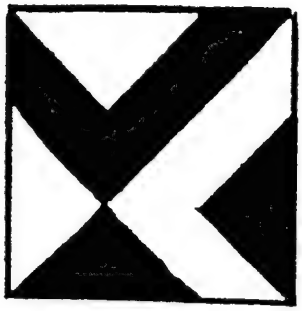
49



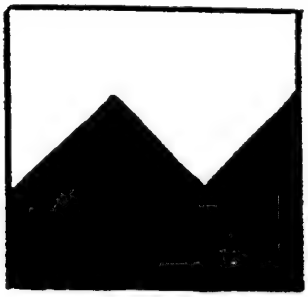
50



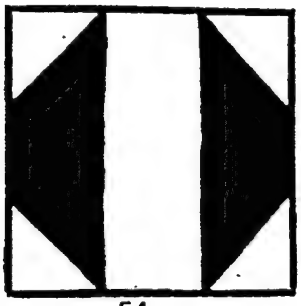
51



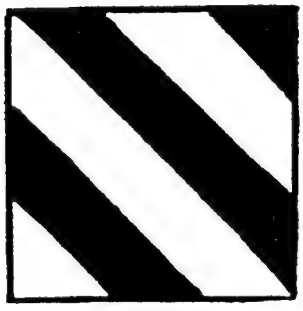
52



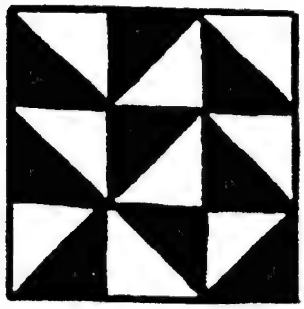
53



54



55



56

Figure 1. (continued)

BLOCK PATTERN CODES

4 = 4 block pattern
9 = 9 block pattern
VE = very easy pattern
E = easy pattern
M = moderate pattern
D = difficult pattern

W = WAIS-R original pattern
C = color reversal of a WAIS-R pattern
R = rotated WAIS-R pattern
X = pattern created by experimenter
P = pattern used for practice only

1.) 4-E-C	13.) 4-E-R	25.) 4-E-X	37.) 9-H-X-P	49.) 9-M-X-P
2.) 4-E-C	14.) 4-E-X	26.) 4-E-X	38.) 9-M-X	50.) 9-M-X
3.) 4-YE-X	15.) 4-E-R	27.) 4-YE-X	39.) 9-M-X	51.) 9-M-X
4.) 4-E-X	16.) 4-E-X-P	28.) 9-M-X	40.) 9-M-X	52.) 9-H-W
5.) 4-E-X	17.) 4-YE-C	29.) 9-M-X	41.) 9-M-X	53.) 9-M-X
6.) 4-YE-X	18.) 4-VE-W	30.) 9-M-X	42.) 9-M-X	54.) 9-M-X
7.) 4-YE-X	19.) 4-E-R	31.) 9-H-R	43.) 9-H-R-P	55.) 9-H-R
8.) 4-VE-W	20.) 4-VE-R-P	32.) 9-H-R	44.) 9-M-X-P	56.) 9-M-X
9.) 4-E-X	21.) 4-VE-X-P	33.) 9-H-R	45.) 9-M-W	
10.) 4-VE-R	22.) 4-YE-X	34.) 9-M-X	46.) 9-H-X-P	
11.) 4-YE-X	23.) 4-VE-R	35.) 9-M-R	47.) 9-M-X	
12.) 4-YE-C	24.) 4-E-W	36.) 9-M-X	48.) 9-M-X	

Figure 1. (continued)

Results

Eight independent observers were asked to provide magnitude estimations and rank orderings of the 56 puzzle patterns. Three types of statistical analyses were conducted: correlations, descriptive statistics, and regression. First, correlations were computed to assess the reliability of these judgments within raters (comparing magnitude estimation to ranking) and between raters on each scaling method. See Table 2 showing stem-and-leaf plots of these three reliability distributions. Mean inter-rater reliability in the range of $r = .80$ and above is considered to be satisfactory for testing instruments (Guilford, 1956).

- The intra-rater reliability between magnitude estimations and rank orderings was assessed using the Spearman's ρ because the rank orderings are ordinal data; ρ ranged from .88 to .97 with a median of .94.

- Inter-rater reliability for the magnitude estimations was assessed using Pearson's r. The r ranged from .70 to .97 with a mean of .88.

- Inter-rater reliability for the rank orderings was assessed using Spearman's ρ; ρ ranged from .73 to .96 with a median of .87.

Table 2

Stem-and-Leaf Plots of Intra- and Inter-Rater Reliabilities

(Stem and leaf plots are a method of displaying frequency distributions in a summary form while still retaining the individual data values. For example, the individual r values for the intra-rater reliabilities are .88, .89, .92, .94, .95, .97, .97, .97.)

Intra-rater reliabilities

.8 8 9
.9 2 4 5 7 7 7

Inter-rater reliabilities (magnitude estimations)

.7 0 9 9
.8 0 0 4 4 4 7 8 8 8 8 8 9 9
.9 0 1 1 1 2 3 4 4 5 5 5 7

Inter-rater reliabilities (rank orderings)

.7 3 6 7 8 8
.8 0 1 3 4 5 5 5 6 7 7 7 7 8 8 8 9
.9 0 0 1 1 2 4 5

Second, the mean and standard deviations (SDs) of the magnitude estimation for each card were calculated to define the pattern's difficulty level. Magnitude estimations were used because they are interval data, while ranks are only ordinal. When two measures have similar reliabilities, the interval measure allows more powerful statistical manipulations (e.g., mean instead of median). Table 3 provides the individual magnitude estimations from each observer and the mean magnitude estimation. Table 4 provides the individual rank orderings and the median rank ordering.

Table 3

Magnitude Estimations for Individual Raters and Mean Estimation for Each Puzzle Pattern Card

CARD	MAG 1	MAG 2	MAG 3	MAG 4	MAG 5	MAG 6	MAG7	MAG 8	AVG MAG	STDEV
1	50	40	30	45	2	40	30	45	38.13	8.84
2	3	7	20	40	40	25	25	25	23.13	13.39
3	10	3	2	15	5	20	10	1	8.25	6.76
4	48	45	35	40	45	40	35	43	41.38	4.75
5	50	20	35	40	50	35	45	38	39.13	9.80
6	2	25	20	10	35	5	20	20	17.13	10.91
7	1	1	2	5	1	5	5	7	3.38	2.39
8	25	10	5	10	15	20	20	27	16.50	7.80
9	45	25	25	30	50	40	40	28	35.38	9.62
10	8	2	2	35	32	20	15	13	15.88	12.53
11	7	3	8	15	15	40	5	12	13.13	11.73
12	12	20	15	10	42	15	20	4	17.25	11.30
13	35	35	25	15	48	35	30	40	32.88	9.88
14	30	5	20	40	50	30	20	22	27.13	13.75
15	35	35	20	15	48	25	30	40	31.00	10.80
16	10	5	10	20	25	35	30	21	19.50	10.52
17	25	10	5	20	15	35	10	3	15.30	10.81
18	15	10	10	30	20	25	25	9	18.00	8.14
19	20	30	15	10	42	15	30	36	24.75	11.40
20	12	21	10	15	30	15	25	24	19.00	7.05
21	15	10	5	50	5	25	15	28	19.13	15.04
22	5	2	5	10	18	20	20	11	11.38	7.21
23	10	10	10	10	30	30	25	10	16.88	9.61
24	40	30	15	7	40	20	30	37	27.38	12.24
25	5	5	20	35	26	40	30	30	23.88	13.04
26	40	15	30	30	38	40	40	17	31.25	10.32
27	1	1	5	5	1	10	1	14	4.75	4.92
28	95	60	75	80	82	75	65	59	73.88	12.22
29	58	58	75	85	55	80	75	76	70.25	11.49
30	62	60	80	60	85	75	75	66	70.38	9.66
31	80	67	90	90	70	85	90	88	82.50	9.32
32	96	75	85	75	70	85	70	81	79.63	8.91
33	63	90	90	60	88	75	90	74	78.75	12.54
34	60	60	75	70	60	80	95	84	73.00	12.95
35	57	55	70	55	60	80	60	64	62.63	8.62
36	72	55	75	60	52	70	75	85	68.00	11.31
37	70	57	90	80	75	85	75	84	77.00	10.34
38	84	60	75	90	78	75	75	80	77.13	8.69
39	64	80	90	60	95	65	95	73	77.75	14.34
40	90	58	85	80	53	95	80	94	79.38	15.84
41	92	65	85	80	85	75	85	73	80.00	8.60
42	75	70	85	90	68	85	85	79	79.63	7.96
43	70	70	90	65	60	80	90	91	77.00	12.39
44	90	57	85	70	58	75	80	81	74.50	12.08
45	51	55	70	65	65	75	60	62	62.88	7.74
46	73	65	85	60	75	85	85	83	76.38	9.84
47	86	65	80	70	75	85	85	82	78.50	7.76
48	98	65	85	100	55	95	80	75	81.63	16.16
49	89	60	70	55	100	85	80	57	74.50	16.55
50	55	60	80	65	80	70	80	70	70.00	9.64
51	65	60	70	75	90	65	80	77	72.75	9.74
52	100	75	85	65	85	85	70	82	80.88	10.84
53	80	60	80	51	53	75	75	78	69.00	12.20
54	56	55	75	55	75	70	70	68	65.50	8.77
55	62	80	90	51	96	75	90	78	78.00	15.52
56	59	52	75	100	55	75	60	53	66.13	16.43

Table 4

Rank Orderings for Individual Raters and Average (MDN and mean) for Each Puzzle Pattern Card

CARD	RANK 1	RANK 2	RANK 3	RANK 4	RANK 5	RANK 6	RANK 7	RANK 8	MDN RANK	Mean RANK
1	20	23	24	23	17	14	25	27	23.00	21.63
2	14	22	12	7	19	13	10	15	13.50	14.00
3	3	5	3	3	3	10	2	1	3.00	3.75
4	25	27	27	24	18	12	27	23	24.50	22.88
5	27	25	15	27	27	27	26	20	26.50	24.25
6	16	2	23	9	2	2	18	11	10.00	10.38
7	2	1	2	1	1	1	1	4	1.00	1.63
8	10	18	18	11	10	11	8	17	11.00	12.88
9	23	19	22	26	24	23	22	12	22.50	21.38
10	7	9	5	17	20	19	5	6	8.00	11.00
11	4	7	4	4	15	25	4	10	5.50	9.13
12	15	6	20	13	4	4	16	3	9.50	10.13
13	22	21	25	21	16	15	23	26	21.50	21.13
14	11	20	8	25	11	24	17	13	15.00	16.13
15	21	26	26	22	21	8	24	25	23.00	21.63
16	17	24	9	19	23	26	15	18	18.50	18.88
17	5	16	19	16	6	5	9	2	7.50	9.75
18	12	10	11	6	13	16	12	9	11.50	11.13
19	18	13	21	14	8	6	19	21	16.00	15.00
20	9	17	17	10	9	9	7	16	9.50	11.75
21	6	8	7	5	12	21	13	24	10.00	12.00
22	8	4	6	12	22	20	6	5	7.00	10.38
23	13	11	10	8	14	17	11	8	11.00	11.50
24	19	14	16	15	7	7	20	22	15.50	15.00
25	24	12	13	18	25	22	14	14	16.00	17.75
26	26	15	14	20	26	18	21	19	19.50	19.88
27	1	3	1	2	5	3	3	7	3.00	3.13
28	39	47	37	41	39	53	36	30	39.00	40.25
29	38	38	32	35	45	52	34	44	38.00	39.75
30	35	37	43	54	41	33	47	41	41.00	41.38
31	49	44	50	40	47	54	53	54	49.50	48.88
32	42	48	56	50	43	49	50	48	48.50	48.25
33	54	49	53	44	30	42	56	39	46.50	45.88
34	56	36	39	49	42	31	39	52	40.50	43.00
35	28	31	29	32	37	43	28	31	31.00	32.38
36	37	43	33	29	33	51	30	49	35.00	38.13
37	34	46	47	34	54	36	40	53	43.00	43.00
38	44	35	34	46	53	38	33	42	40.00	40.63
39	55	32	52	43	29	29	54	33	38.00	40.88
40	41	53	46	55	55	50	48	56	51.50	50.50
41	45	50	44	52	40	47	38	37	44.50	44.13
42	47	40	41	42	52	45	51	35	43.50	44.13
43	48	51	51	53	46	46	52	55	51.00	50.25
44	33	29	35	39	48	34	35	46	35.00	37.38
45	29	30	28	31	38	30	29	38	30.00	31.63
46	52	54	40	47	50	37	44	51	48.50	46.88
47	51	52	48	37	51	32	45	43	46.50	44.88
48	40	55	42	51	56	56	41	50	50.50	48.88
49	50	45	49	56	49	44	43	29	47.00	45.63
50	32	34	45	36	35	35	37	34	35.00	36.00
51	46	39	36	38	32	41	42	32	38.50	38.25
52	43	56	55	48	44	48	49	47	48.00	48.75
53	31	41	38	33	28	28	46	45	35.50	36.25
54	36	42	31	28	36	40	31	36	36.00	35.00
55	53	33	54	45	31	39	55	40	42.50	43.75
56	30	28	30	30	34	55	32	28	30.00	33.38

Finally, a regression analysis was calculated to provide a visual representation of the reliability and the mean trend for the assessed difficulty of these 56 patterns. The analysis regressed the individual magnitude estimations (as the dependent variable) against the mean magnitude estimation (as the independent variable). First, regression was computed on the entire set of 56 cards (including both the four- and the nine-block patterns). The linear regression equation was $Y' = 1.64 X + 4.12$ and the $R^2 = .82$ or explaining 82% of the variance (see Figure 2).

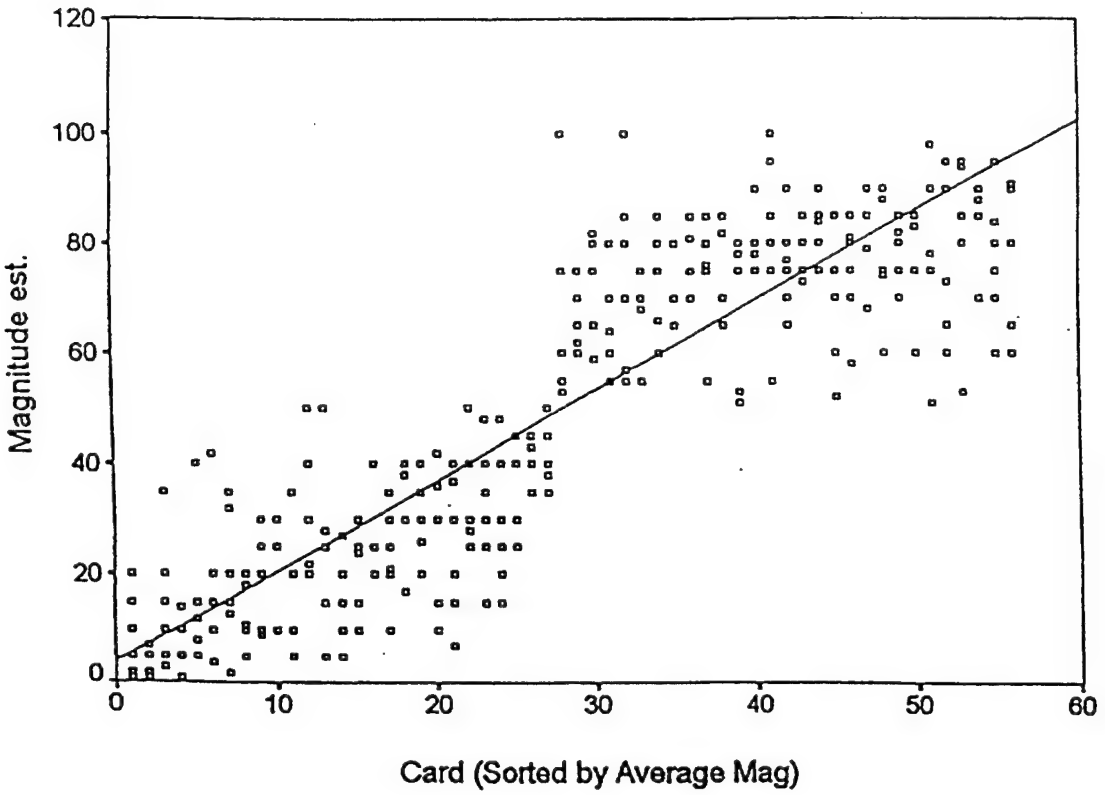


Figure 2. Regression equation and scatter plot showing magnitude estimates for all 56 cards.

A portion of the strong correlation reflects the clear separation between the assessed difficulty of the four-block and the nine-block patterns. This is consistent with the information obtained from the WAIS-R manual and with the manner in which magnitude estimations were assigned (1 to 50 for four-block and 51 to 100 for nine-block). Clearly, two levels of difficulty exist in the total set of 56 patterns.

Next, the regression was calculated within each of the two pattern sets (four- and nine-block patterns):

• The regression equation for the four-block patterns was Y' = 1.14 X + 7.24 and the R2 = .49 or explaining 49% of the variance. The F-test of the significance of the explained variance (greater than using the set of four-block patterns as a single undifferentiated difficulty level) is F = 182.8, p < .0l (see Figure 3).

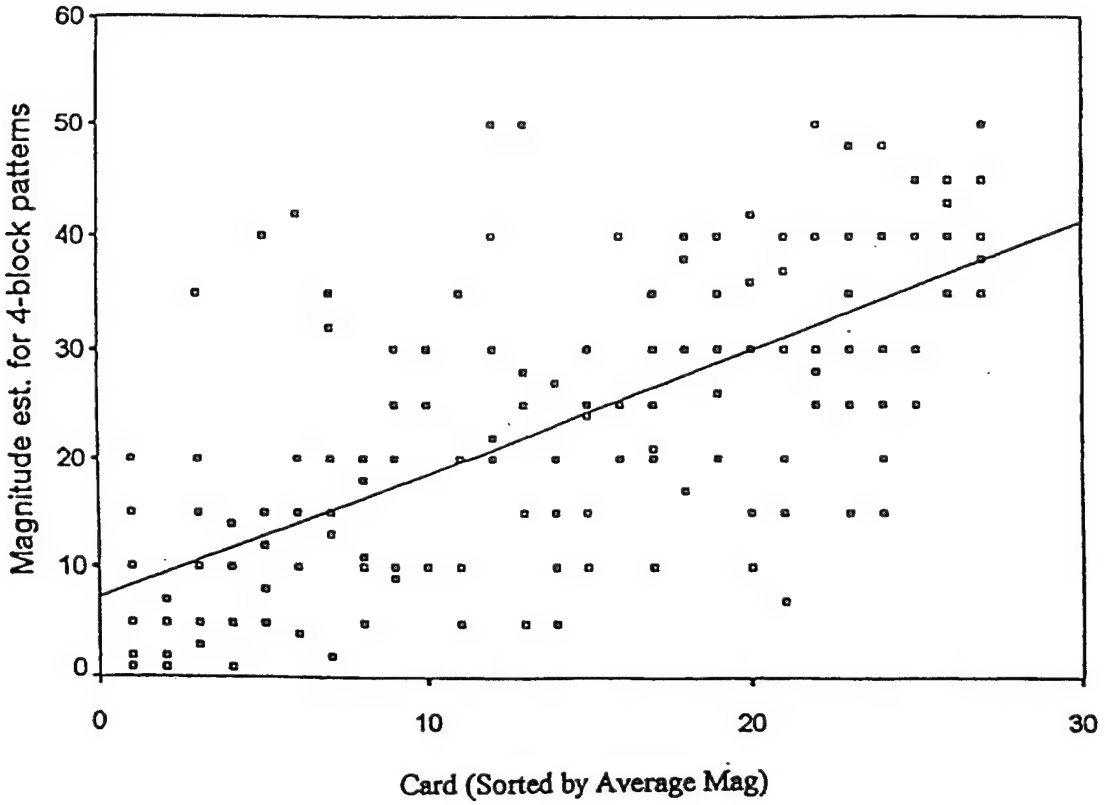


Figure 3. Regression equation and scatter plot showing magnitude estimates for the four-card patterns.

• The regression equation for the nine-block patterns was Y' = .58 X + 52.32 and the R2 = .20 or explaining 20% of the variance. The F-test of the significance of the explained variance (greater than using the set of nine-block patterns as a single undifferentiated difficulty level) is F = 59.9, p< .01 (see Figure 4).

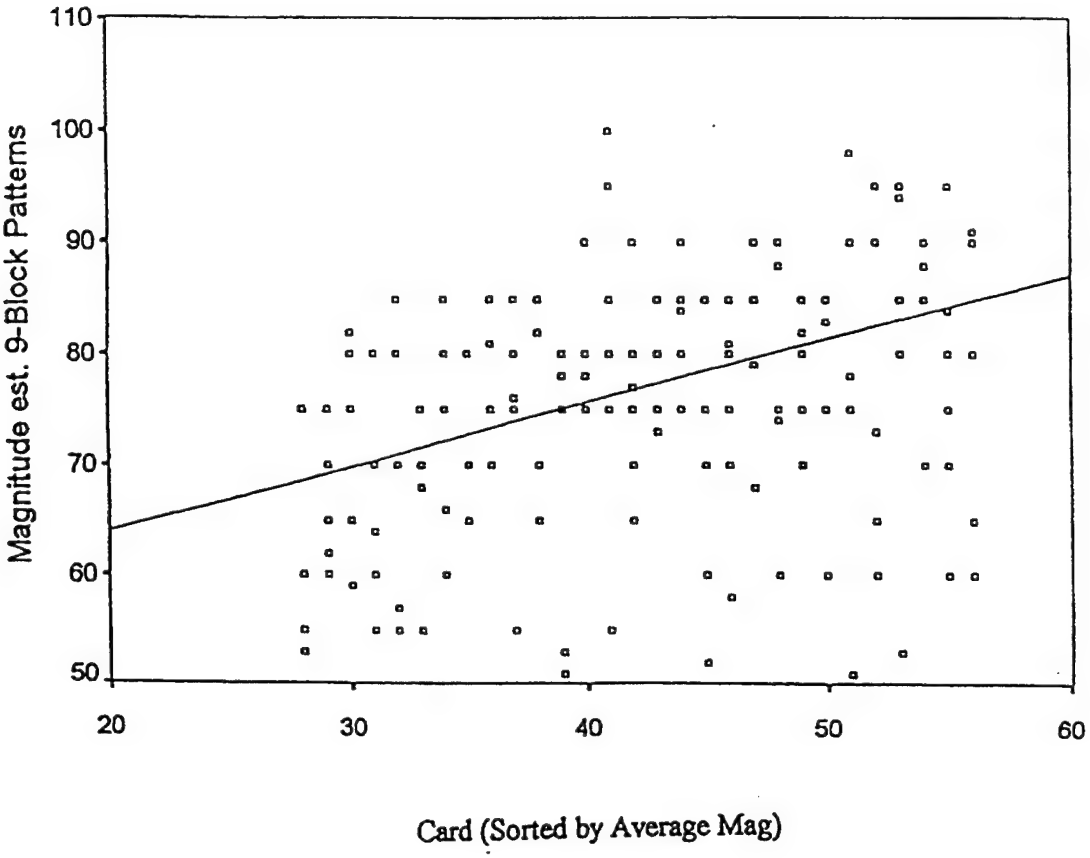


Figure 4. Regression equation and scatter plot showing magnitude estimates for the nine-card patterns.

The significant F value for each regression analysis indicates that there is a reliable change in the difficulty level within both the four-block and the nine-block patterns as well as between the two pattern sets. These results are consistent with using the 56 patterns at four separate levels of difficulty. There may be more separable difficulty levels, but four will be a practical number to test the workload metrics in the proposed research. A quartile split was used to define four pattern difficulty levels: very easy, easy, moderate, and difficult.

Assessing Spatial Ability: Cognitive Laterality Battery (CLB)

Rationale for Measuring Spatial Ability

Individual differences in subject's ability to perform various tasks can possibly cloud the results obtained in experimental research. Therefore, investigators either control this possibility by holding individual difference variables constant or by measuring such variables and stratifying the sample to allow the measurement of their impact. In the present study, two of these variables might be verbal intelligence and spatial ability. Intelligence within either a sample of university students or a sample of medical personnel is not likely to vary greatly. Selection into these populations has already greatly restricted the range since verbal intelligence is highly correlated with academic success. However, spatial ability may range widely within either population because it is not so directly correlated with any selection procedure for academic success. Hence, a measure of spatial ability was sought with which to stratify the subjects in our proposed experiment. By this means, it would be possible to determine whether spatial ability was a variable affecting performance of the task in general, or interacting with either task difficulty level or communication method (co-location versus telemedicine).

Selecting Spatial Ability Instrument

A spatial ability test battery called the *Cognitive Laterality Battery* has been developed, validated, and normed by Gordon (1987). The entire test is a cognitive laterality battery intended to determine the specialized functioning in each cerebral hemisphere. Of interest for the present research are the four subscales (tests) that comprise the measurement of spatial ability. These four tests are called localization, orientation, form completion, and touching blocks. Each measures some aspect of spatial ability, and collectively, they provide a reliable measure of this ability.

Cognitive Laterality Battery Test Materials

The CLB is available commercially in a package that includes all administration instructions, stimulus materials in the form of slides and taped instructions, data sheet templates, and scoring instructions and answer keys. In addition, the norming data (means, standard deviation, and frequency distributions) for several populations are provided.

Equipment to Administer the Cognitive Laterality Battery

To administer the four spatial ability tests, the following equipment is used: slide

projector (Kodak Carousel 5400) and tape recorder-player (General Electric #3-5622A). The administration of the tests, including their instructions and material distribution, requires approximately 60 minutes; of this time, 30 minutes are required for actual data collection (time spent viewing the stimuli and marking responses). Subjects can be tested in groups of as many as 10 people, depending upon the viewing conditions. It is necessary for each subject to be able to see clearly the stimulus slides projected on a screen.

Spatial Ability Subscales

The four spatial ability subscales (i.e., *localization, orientation, form completion, and touching blocks)* are described next:

- *Localization* is a test of the observer's ability to reproduce the location of an x marked on a projected slide by marking its corresponding location on a paper template. There are 24 slides.
- *Orientation* is a mental rotation task. Observers view three 3D geometric figures and determine which two figures are actually the same object. There are 24 tasks.
- *Form Completion* consists of line drawings of common figures with portions of the line segments erased (missing). The observer's task is to name the figure. There are 24 figures.
- *Touching Blocks* shows a stack of blocks in which some blocks are numbered. The observer's task is to count the number of blocks touching all the numbered blocks. There are six stacks.

Code Results

The results are scored by referring to the answer key for each test, except the location subscale. The location subscale requires that the experimenter score the distance in millimeters that the observer's response is from the target location. This is a time-consuming scoring procedure, even using the template provided in the test booklet.

Tabulate Results

The results can be used as subscale values so that they can be compared to the adult norming values for each subscale in the CLB manual. Alternatively, a general spatial abilities score can be obtained by adding all the subscale scores for a given subject. The score for the localization subscale is an error measurement and hence is negatively correlated with spatial ability. Therefore, the actual localization score can be subtracted from any constant larger

than the largest error score in the sample. This transformed score will then be positively correlated with spatial ability and can be added to the remaining three subscale scores to obtain a total spatial ability measure for each subject.

Application for Proposed Testing

The four spatial ability subscales of the Cognitive Laterality Battery are available, reliable, validated, uncontaminated, and manageable methods of measuring spatial ability. The CLB is recommended as a satisfactory method of stratifying spatial ability.

Developing Test Paradigm Procedures

The final activity in completing testing of this paradigm was to design and test the research protocol itself. A generic workload measure was sought, which will assess the cognitive requirements that are likely to exist in most medical procedures. Furthermore, there is a specific interest in targeting the changes in cognitive workload that occur with the introduction of telecommunication for those procedures. Hence, a research protocol to test the interaction of three variables was designed. The three variables are *type of workload metric, task difficulty, and communication condition.* A measure for stratifying subjects by *spatial ability* was included.

Workload Metric

The three candidate workload measures selected were the **Subjective Workload Assessment Technique (SWAT)** (Reid & Nygren, 1988), the **NASA-Task Load Index (TLX)** (Hart & Mashkati, 1988), and the **Modified Cooper-Harper (MCH)** (Boff & Lincoln, 1988).

Task Difficulty

A surrogate puzzle pattern task was developed as described earlier. The magnitude estimations of difficulty were used to produce four separate levels of task difficulty which will be used to assess the sensitivity of the three workload metrics. Thirteen patterns of each difficulty level were designed and tested.

Communication Condition

The two communication conditions are co-location and telecommunication. In the co-location condition, the two team members are located in the same room and view the working area directly. In the telecommunication condition, they are located in separate rooms and have to communicate via video and audio communication.

Spatial Ability

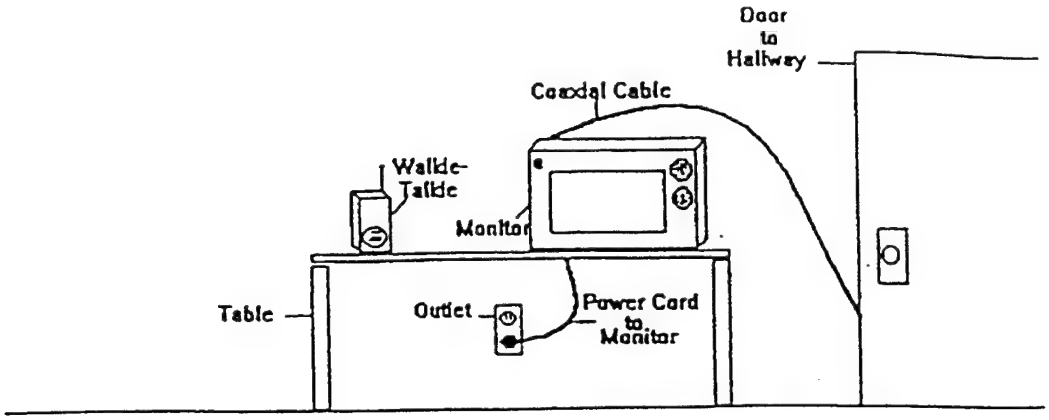
The four types of spatial ability teams are constructed by using the subjects' scores on the spatial ability subscales of the Cognitive Laterality Battery. The four types of teams are high:high, low:high, high:low, low:low, in which the first member is the instructor and the second is the builder.

Equipment

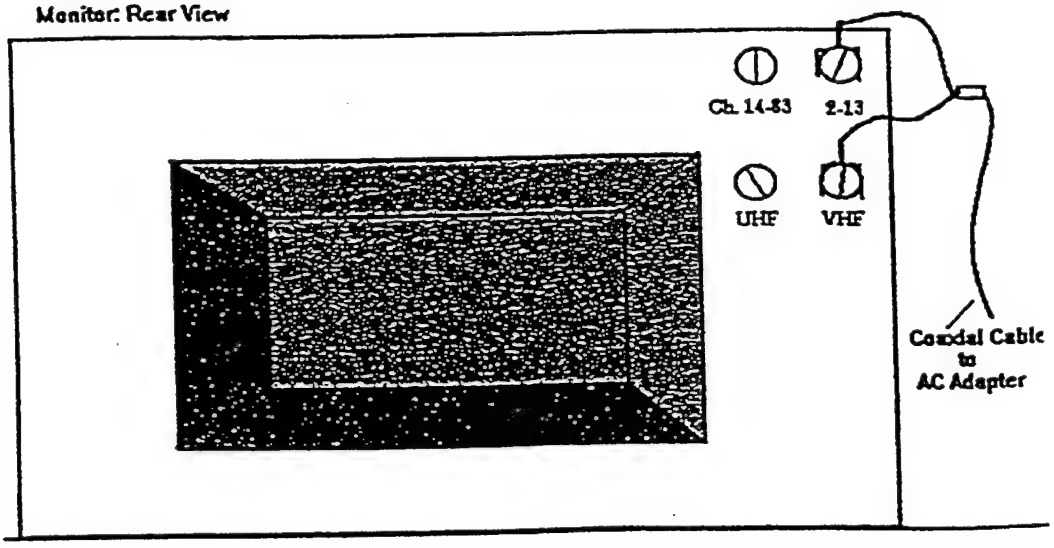
To test the feasibility of the anticipated experiment, it was necessary to determine the telecommunication equipment that would be used in the experiment. The major video components of this equipment were obtained as a loan from the U.S. Army Research Laboratory at Aberdeen Proving Ground, Maryland. These consisted of two video cameras (Panasonic VHS AG 160 Proline camcorder and AC adapter) and two television monitors (19-inch Zenith Model No. L1912W). Additional equipment, which was obtained from local sources, consisted of two TRC-512, 49-MHz FM Radio Shack wireless transmitter-receivers ("walkie-talkies") to permit audio communication between the team members in the telemedicine condition, a RST-84V Radio Shack tripod, and a 25-foot coaxial cable to connect the remote monitor to the camcorder. See Figure 5 for diagram of the equipment setup.

Design

To select the best workload metric for use in evaluating telemedicine applications, the following mixed factors design with three independent variables was developed. Spatial ability of teams is varied at four levels: high:high, high:low, low:high, and low:low. The remaining variables are both repeated measures: communication condition (co-located versus telecommunication) and four levels of puzzle pattern difficulty (very easy, easy, moderate, and difficult). Each level of puzzle difficulty occurs on a total of 12 trials. Half of these are in the co-located and half in the telecommunication condition. In each communication condition, workload for two of the trials is assessed using each of the three workload metrics (SWAT, NASA-TLX, and MCH). Thus, a total of 48 trials (puzzle patterns) are completed by each team. A diagram of this mixed factors design is as follows: 4 spatial ability x (2 communication x 4 task difficulty x 3 workload metrics x 2 replications). All levels of the repeated measures variables will be counterbalanced or randomized to avoid confounding order with experimental treatment results.

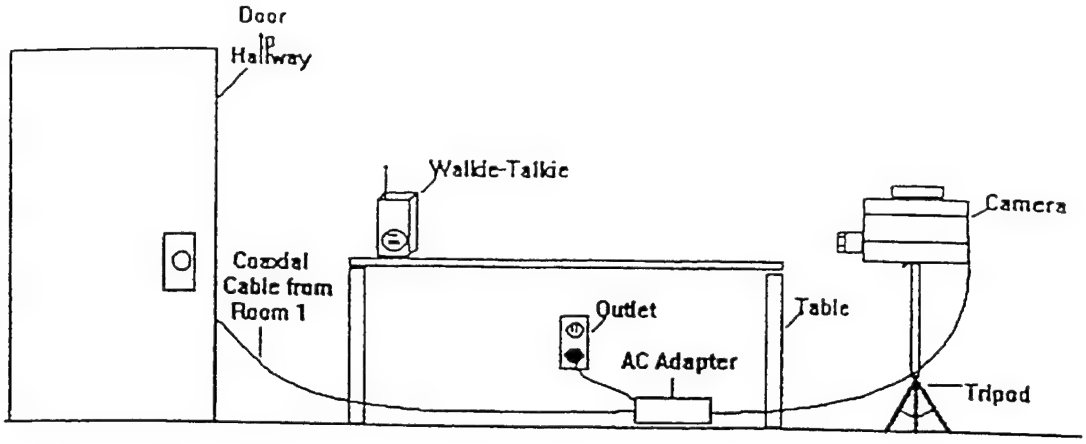


Room 1 (for telecommunication)

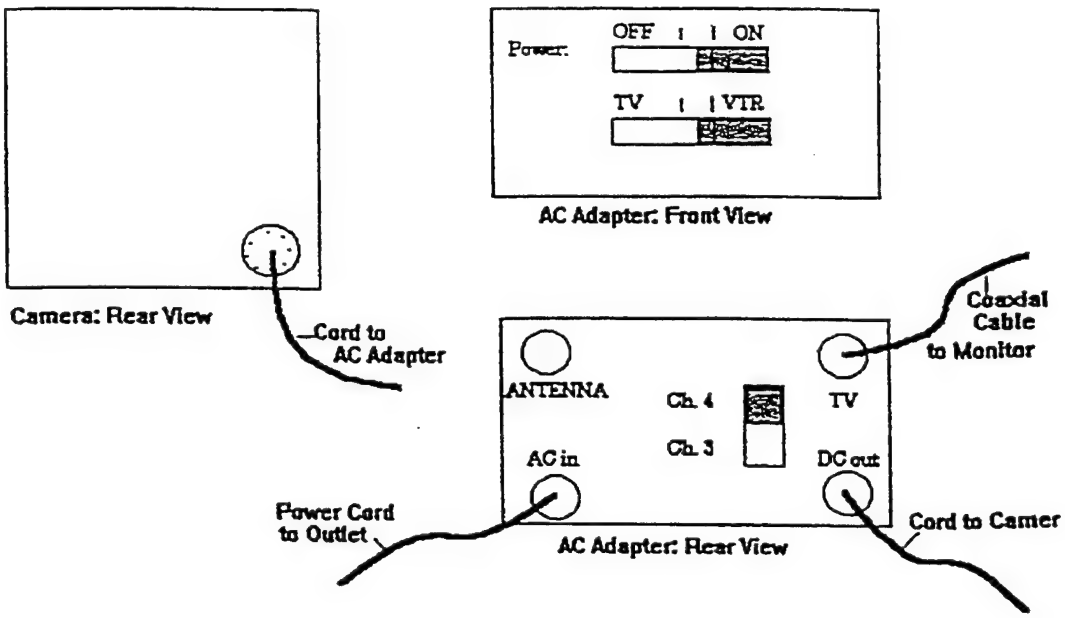


Monitor Connections

Figure 5. Diagrams of equipment setup.



Room 2 (for telecommunication)



Camcorder Connections

Figure 5. (continued).

Test Procedure and Modify Iteratively

The actual procedure for the experimental paradigm required modification from its conceptualization to its final form. This iteration was accomplished by the principal investigator and the research assistant alternatively serving as experimenter and subject or both as team members until the *procedure, instructions, equipment, training*, and *measurement* issues had been satisfactorily developed. The following parameters were established empirically during these iterative modifications:

- Number of training trials
- Preview time for patterns
- Audio communication equipment
- Field of view and camera angle
- Permissible puzzle patterns constrained by video view
- Instructions to team members
- Method of recording errors (sketch)
- Anticipated number of errors influenced design of dependent variables.
- Time allowed for pattern building
- Power for obtaining workload measures (by two-trial blocks, not for each trial.)

Procedure

Subjects are introduced to the experimental room and the communication equipment. They are told that their task is to work together as teams to build a series of puzzle patterns from blocks. Before data collection begins, each team completes seven practice trials in which they become familiar with one another's terminology and typical strategies.

In the t*elecommunication condition,* one team member, serving as the instructor, sits in the room with the television monitor (Room I) and the other, serving as the builder, in the room with the camcorder (Room B). The instructor has a stack of 24 patterns. The instructor's task is to describe how to build a given pattern. The builder has the blocks on the table. The builder's task is to build the pattern that is described. Both subjects view the two patterns for a given condition (e.g., moderate difficulty, telecommunication, MCH) for 10 seconds. After the preview, the instructor is the only one to see the paper pattern. The measure of time begins when the experimenter says "begin" for each trial. It ends when the instructor signals completion. After both trials are completed in a given condition, a workload rating is obtained.

A similar procedure is used for the *co-location condition* except that the two team members are in the same room. Again, the builder is the only team member allowed to touch the blocks and the instructor is the only one to see the paper pattern. Time to completion, errors in

pattern built (including the sketch of any incorrect result), and workload ratings are recorded as the dependent variables.

The instructions for the instructor and the builder team members in both the telecommunication and the co-location conditions are given in Table 5.

Complete Protocol With a Sample Team

After the procedural modifications are completed, the entire protocol (omitting the CLB) was completed using two graduate students as subjects. The procedure required 2 hours to complete all seven training trials and 48 data collection trials. The results for this team are summarized in Table 6.

On the basis of these data, two further changes were made in the protocol:

- The experimenter's procedure checklist was changed to make it easier to collect the results without the errors that led to the loss of the NASA-TLX data in the sample run.
- Debriefing questions were added to collect information systematically about the subject's preferences for one or another of the workload measures.

Conclusions

The work described in this paper was undertaken to establish a research paradigm for developing a satisfactory evaluation tool for telemedicine applications. These efforts were successful in establishing the feasibility of that research. A surrogate task (team building of block patterns) was developed and 56 patterns of measured difficulty were designed, produced, and tested. This task can be used as it is and can be modified to incorporate a greater psychomotor component, when such a component proves necessary in some experiments. The telecommunication equipment necessary to conduct the research was acquired, set up, and tested. The possibility of uncontrolled individual differences in spatial ability was considered for some populations and a measure of spatial ability was determined so that teams can be stratified on this measure. A scientifically sound research design and a procedure for implementing that design were developed. The entire procedure was tested and final adjustments were made.

Table 5

Instructions for Builder and Instructor in Telecommunication and Collocation Conditions

Telecommunication Condition:

Instructions for the builder: In this portion of the study you will receive a set of instructions given to you by your team mate located in another room. You will hear these instructions over your walkie-talkie. You will place these red and white blocks as you are told to form one of the three patterns you have viewed. The blocks consist of two red sides, two white sides, and two sides split in half so that they are both red and white. This camera is here so that your team mate may monitor your progress and correct any mistakes you may make. Some patterns will seem harder than others. After completing three patterns, you will be asked to fill out a form that describes the amount of work you think was involved in completing the previously built patterns. This is a subjective measure and will not be the same for all people so do not feel as though your ratings must meet a set standard. After you have completed the measure of workload, you will build three more patterns and fill out another workload evaluation and so on until all patterns are completed (there are twenty-seven). Your goal is to work as quickly as possible while attempting to build a complete correct pattern. Your team will receive a twenty-five-dollar reward if it is one of the two fastest teams with the fewest errors. You may ask your team mate to repeat any instructions you do not understand by depressing the talk button on your own walkie-talkie. Are there any questions?

Instructions for the person with the patterns: In this portion of the study you will be asked to describe these patterns you see before you now to your team mate located in another room. Your team mate has a set of blocks in order to achieve this construction which have two red sides, two white sides, and two sides that are split in half so that they are both red and white. You will communicate to your team mate via a set of walkie-talkies, one of which you see before you. You talk by depressing the talk button for the duration of the time you need to speak. Your team mate has the option of asking you to repeat any instructions s/he does not understand. Keep in mind your team mate has viewed the patterns you are describing for thirty seconds for nine-block patterns and fifteen seconds for four-block patterns. The television monitor is here so that you may monitor your team mate's progress and correct any errors s/he may make. Once you have explained three patterns you will be asked to fill out a workload evaluation which will let the experimenter know how much work you believe was involved in completing this phase of the experiment. When this evaluation is completed, you will describe three more patterns and receive another evaluation and so on until all patterns are completed (there are twenty-seven). Workload evaluations are subjective; therefore your opinions may or may not match someone else's. Do not worry; you are not trying to meet a standard, just state your own opinion. Your goal is to complete these patterns as quickly as possible, making as few errors as possible. At the end of the experiment, the two teams with the fastest times and the fewest errors will receive a twenty-five-dollar bonus. Are there any questions?

Collocated Condition:

Instructions for both subjects: In this phase of the study you will be asked to construct the patterns you see before you. Only one of you will have access to the patterns while the other will have the blocks. However, you both will be permitted to view the three patterns occurring in the ensuing block. If the patterns contain nine blocks, you will be allowed to view them for thirty seconds and if there are four blocks, you may view them for fifteen seconds. Only one designated person may touch the blocks. The person with the designs must describe to the other person how to situate the blocks in order to create the pattern s/he sees. Each block consists of two red sides, two white sides, and two sides split in half so that they are both red and white. The builder may at any time ask the instructor to repeat instructions that were not understood; however, the builder may not ask to see the design itself nor may the instructor show the design to his or her team mate. After completing three designs, you will both be asked to fill out a workload evaluation which will tell the experimenter how much work you each feel was involved in completing this phase of the experiment. These evaluations are subjective so the evaluations you both fill out may not reflect the same ideas. Do not worry about matching your partner's evaluation; the experimenter wants to know what each of your personal views are. When this evaluation is completed, you will be asked to complete three more patterns and give another evaluation and so on until all patterns are completed (there are twenty-seven). Your goal is to complete the patterns as quickly as possible while making as few errors as possible. At the end of the experiment, the two teams with the fastest times and the fewest errors will receive a twenty-five-dollar reward. Are there any questions?

Table 6

Sample of One Team's Performance of Experimental Protocol

		Collocation		
Pattern difficulty	Very easy	Easy	Moderate	Difficult
Time (in sec.)	9	9.8	24	25.8
SWAT	33	33	72	72
NASA-TLX	Lost: Experimenter error			
MCH	30	20	50	60
		Telecommunication		
Pattern difficulty	Very easy	Easy	Moderate	Difficult
Time (in sec.)	15.8	27.67	39.17	35.17
SWAT	33	44	61	61
NASA-TLX	Lost: Experimenter error			
MCH	20	30	30	70

NOTE: WL adjusted to 0 to 100 range

REFERENCES

Acton, W.H., Crabtree, M.S., Simons, J.C., Gomer, F.E., & Eckel, J.S. (1983). Quantification of crew workload imposed by communications-related tasks in commercial transport aircraft. Proceedings of the Human Factors Society 27th Annual Meeting, 239-243.

Bashshur, R.L. (1995). On the definition and evaluation of telemedicine. Telemedicine Journal, 1, 19-30.

Biers, D.W., & McInerney, P. (1988). An alternative to measuring subjective workload: Use of SWAT without the card sort. Proceedings of the 32nd Annual Meeting of the Human Factors Society, 1136-1139.

Boff, K.R., & Lincoln, J.E. (1988). Engineering data compendium: Human perception and performance. Dayton, OH: Armstrong Aeromedical Research Laboratory, Wright-Patterson AFB.

Boyd, S.P. (1983). Assessing the validity of SWAT as a workload measurement instrument. Proceedings of the Human Factors Society 27th Annual Meeting, 124-128.

Crabtree, M.S., Bateman, R.P., & Acton, W.H. (1984). Benefits of using objective and subjective workload measures. Proceedings of the Human Factors Society 28th Annual Meeting, 950-953.

Derrick, W.L. (1983). Examination of workload measures with subjective task clusters. Proceedings of the Human Factors Society 27th Annual Meeting, 134-138.

Detro, S.D. (1985). Subjective assessment of pilot workload in the advanced fighter cockpit. Paper presented at the Third Symposium on Aviation Psychology, Ohio State University, Columbus, OH.

Eggemeier, F.T., Crabtree, M., & LaPointe, P. (1983). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society 27th Annual Meeting, 139-143.

Eggemeier, F.T., McGhee, J.Z., & Reid, G.B. (1983). The effects of variations in task loading on subjective workload rating scales. Proceedings of the IEEE 1983 National Aerospace and Electronics Conference, 1099-1105.

Eggemeier, F.T., Melville, B.E., & Crabtree, M.S. (1984). The effect of intervening task performance on subjective workload ratings. Proceedings of the Human Factors Society 28th Annual Meeting, 954-958.

Eggemeier, F.T., & Stadler, M.A. (1984). Subjective workload assessment in a spatial memory task. Proceedings of the Human Factors Society 28th Annual Meeting, 680-684.

Eggleston, R.G. (1984). A comparison of projected and measured workload ratings using the subjective workload assessment technique (SWAT). Proceedings of the National Aerospace and Electronics Conference, 827-832.

Fisk, A.D., Derrick, W.L., & Schneider, W. (1983). The assessment of workload: Dual task methodology. Proceedings of the Human Factors Society 27th Annual Meeting, 229-233.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), Handbook of perception and human performance (Vol. 2). Cognitive Processes and Performance (pp. 41.1-41.9). New York: John Wiley and Sons.

Gordon, H.W. (1987). The cognitive laterality battery. Pittsburgh, PA: University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic.

Grigsby, J., Schlenker, R.E., Kaehny, M.M, Shaughnessy, P.W., & Sandberg, E.J. (1995). Analytic framework for evaluation of telemedicine. Telemedicine Journal, 1, 31-39.

Guilford, J.P. (1956). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hart, S.A., & Mashkati, N. (1988). Development of NASA-TLX (Task Loading Index): Results of empirical and theoretical research. In P. A. Hancock & N. Mashkati (Eds.), Human Mental Workload (pp. 239-250). Amsterdam: North Holland.

Hassel, L.H. (1995, August). Telemedicine evaluation of Project AKAMAI. Paper presented at the Evaluation Methodologies Conference, Ko Olina, Hawaii.

Heffley, R.K. (1983). Pilot workload factors in the total pilot-vehicle-task system. Proceedings of the 27th Annual Meeting of the Human Factors Society, 234-238.

Hendy, K.C., Hamilton, K.M., & Landry, L.N. (1993). Measuring subjective workload: When is one scale better than many? Human Factors, 35, 579-601.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A., Zaklad, A.L., & Christ, R.E. (1992). Comparison of four subjective workload rating scales. Human Factors, 34, 429-439.

Kling, J.W., & Riggs, L.A. (1972). Woodworth & Schlosberg's Experimental Psychology (Vol. 1): Sensation and Perception. New York: Holt, Rindhart & Winston, Inc.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., & Wherry, R.J. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies (Technical Report 851). Ft. Bliss, TX: U.S. Army Research Institute Field Unit.

Moray, N. (1982). Subjective mental workload. Human Factors, 24, 25-40.

Moroney, W.F., Biers, D.W., & Eggemeier, F.T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. International Journal of Aviation Psychology, 5, 87-106.

Nygren, T.E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Human Factors, 32, 17-33.

Philipp, U., Reiche, D., & Kirchner, J.H. (1971). The use of subjective ratings. Ergonomics, 14, 611-616.

Polzella, D.J., & Reid, G.B. (1987). A multidimensional scaling analysis of subjective workload assessment technique (SWAT) ratings of the criterion task set (CTS). Proceedings of the 31st Human Factors Society Annual Meeting, 398-401.

Puskin, D.S., Brink, L.H., Mintzer, P.P.A., & Wasem, C.L. (1995). Joint federal initiative for creating a telemedicine evaluation framework. Telemedicine Journal, 1, 393-397.

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), Human Mental Workload (pp. 185-214). Amsterdam: North Holland.

Schlegel, R.E. (1993). Driver mental workload. In B. Peacock & W. Karwowski (Eds.), Automotive Ergonomics (pp. 359-382). Washington, DC: Taylor and Francis.

Vidulich, M.A., & Tsang, P. (1986). Techniques of subjective workload assessment: A comparison of SWAT and NASA-Bipolar methods. Ergonomics, 29, 1385-1398.

Vidulich, M.A., & Wickens, C.D. (1986). Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments. Applied Ergonomics, 17, 291-296.

Warr, D., Colle, H., & Reid, G. (1986). A comparative evaluation of two subjective workload measures: The subjective workload assessment technique and the modified Cooper-Harper scale. Paper presented at the Symposium on Psychology in the Department of Defense, USAFA, Colorado Springs, CO.

Wechsler, D. (1981). WAIS-R manual: Wechsler adult intelligence scale-revised. Cleveland, OH: Harcourt, Brace, & Jovanovich.

Whitaker, L.A., Peters, L., & Garinther, G. (1989). Tank crew performance: Effects of speech intelligibility on target acquisition and subjective workload assessment. Proceedings of the Human Factors Society 33rd Annual Meeting, 1411-1413.

Wickens, C.D., & Yei-Yu, Y. (1983). The dissociation between subjective workload and performance: A multiple resource approach. Proceedings of the 27th Annual Human Factors Society Meeting, 244-248.

Wierwille, W.W., & Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications. Proceedings of the Human Factors Society 27th Annual Meeting, 129-133.

Wierwille, W.W., & Connor, S.A. (1983). Evaluation of twenty workload assessment measures using a psychomotor task in a motion-base simulation. Human Factors, 35, 263-281.

Wierwille, W.W., & Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. Human Factors, 35, 263-281.

NO. OF COPIES	ORGANIZATION
2	ADMINISTRATOR DEFENSE TECHNICAL INFO CENTER ATTN DTIC DDA 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	DIRECTOR US ARMY RESEARCH LABORATORY ATTN AMSRL CS AL TA RECORDS MANAGEMENT 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LABORATORY ATTN AMSRL CI LL TECHNICAL LIBRARY 2800 POWDER MILL RD ADELPHI MD 207830-1197
1	DIRECTOR US ARMY RESEARCH LABORATORY ATTN AMSRL CS AL TP TECH PUBLISHING BRANCH 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LABORATORY ATTN AMSRL CI LP (TECH LIB) BLDG 305 APG AA
1	LIBRARY ARL BLDG 459 APG-AA

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1997	Final

4. TITLE AND SUBTITLE

Selection of a Workload Metric for Evaluation of Telemedicine Applications: Literature Review and Methodological Development

5. FUNDING NUMBERS

AMS Code 622716.H700011
PR: 1L162716AH70
PE: 6.27.16

6. AUTHOR(S)

Whitaker, L.A.; Hahus, J.; Birkmire-Peters, D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
Human Research & Engineering Directorate
Aberdeen Proving Ground, MD 21005-5425

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
Human Research & Engineering Directorate
Aberdeen Proving Ground, MD 21005-5425

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ARL-TR-1264

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A measure of cognitive workload was needed to conduct human factors evaluations of telemedicine applications. A literature review was conducted to find available metrics and select candidates for testing. Three candidate measures (the Subjective Workload Assessment Technique [SWAT], NASA-Task Load Index [TLX] along with its subscales, and the Modified Copper-Harper [MCH]) were selected using the following criteria: reliability, validity, lack of contamination, availability, sensitivity, lack of intrusiveness, diagnosticity, and cost. All metrics in the literature review, as well as the application of the selection criteria, are described in this report. Methodological development and research were then completed to develop a research paradigm for selecting the best workload metric from the three candidates. This effort included the development and norming of difficulty levels of a surrogate task in a controlled experimental protocol, the selection of a spatial abilities test, acquisition and testing of required telecommunication and recording equipment, and the iterative development and testing of a research protocol. These processes and their results are described in detail in this report.

14. SUBJECT TERMS

telemedicine
workload

15. NUMBER OF PAGES

50

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	